The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Comparing Reward Shaping, Visual Hints, and Curriculum Learning

Rey Pocius,[1] David Isele,[2] Mark Roberts,[3] David W. Aha[3]
[1]Oregon State University, School of EECS, Corvallis, OR 97331, USA | pociusr@oregonstate.edu
[2]University of Pennsylvania, School of EAS, Philadelphia, Pennsylvania, 19104, USA | isele@seas.upenn.edu
[3]Naval Research Laboratory, Code 5514; Washington, DC, 20375, USA | {first.last}@nrl.navy.mil

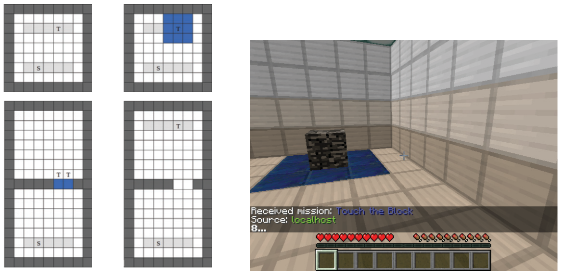
When considering how to reduce the learning effort required for Reinforcement Learning (RL) agents on complex tasks, designers can apply several common approaches. *Reward shaping* boosts the immediate reward provided by the environment, effectively encouraging (or discouraging) specific actions. *Curriculum learning* (Bengio et al. 2009) aims to help an agent learn a complex task by learning a sequence of simpler tasks. *Hints* may also be provided (e.g., a yellow brick road), which fall outside the notion of shaping or a curricula. Despite the prevalence of these approaches, few studies examine how they compare to (or complement) each other or when an approach is better.

As a first step in this direction, we analyze shaping, hints, and curricula for a Deep RL agent in Malmo (Johnson et al. 2016), a research platform for Minecraft. Figure 1 (left) shows the layouts used in our study, which are distinguished by the number of rooms, the placement of the target, and whether color is included. For all rooms, the starting position of the agent is selected from five blocks at the bottom of the room (highlighted gray). In one-room situations and the right-most two-room situation, the target is always chosen from the five blocks at the top of the room (highlighted gray). In the left-most two-room situation, the target is set just beyond the doorway. Visual hints are provided in some situations by coloring some of the floor blocks blue.

We seek to answer whether shaping, hints, or the curricula have the most impact on performance, which we measure as the time to reach the target, the distance from the target, the cumulative reward, or the number of actions taken. For this task, performance is most impacted by the curriculum used and hints while shaping had less impact, suggesting that designing an effective curriculum and providing appropriate hints deserve more attention for similar navigation tasks with Deep RL agents. Our methodology provides an evaluation protocol, serving as a foundation for further studies that tease apart when (and why) methods excel or fail.

Reinforcement Learning (RL)

RL agents learn a mapping of world states to actions (i.e., a policy) through trial and error in dynamic or static environments. Reward signals from the environment guide learning. This type of learning can be done in both discrete and continuous action spaces. Significant advancements in RL have moved the field past discrete action spaces in toy problems toward continuous actions and more challenging domains. As training of agents in these domains have improved through the use of architectures such as Deep Q-Networks (DQNs), researchers have been conducting their research in more realistic simulators (e.g., Malmo, Gazebo).



Received mission: Touch the Block
Source: localhost
8...

Figure 1: Example Minecraft room layouts (left) and a screenshot (right) of the black target and blue floor coloring.

Several challenges come with these kinds of problems. Many of the spaces are very large, making it difficult for agents to fully explore the state-action space during training. Additionally, larger state spaces increase the reward sparsity. Methods such as curriculum learning and shaping have been used to combat these challenges.

Curriculum Learning (Bengio et al. 2009) allows an agent to learn a complex task (e.g., navigating two rooms to a target) by learning a sequence of easier tasks (e.g., navigate in a single room, navigate to a doorway, navigate through two rooms). It leverages the benefits of transfer learning to increase the learning speed and robustness of a target task with the use of many source tasks. Research conducted on curriculum learning has spanned a number of different domains and problems leading to a plethora of curricula design methodologies. Some of these problems require curriculum learning to be solved efficiently. Recently, Narvekar et al. (2016) examined strategies for decomposing a target task into easier source tasks and found that domain-specific curricula design and reward structure choices are needed to achieve optimal behavior. Matiisen et al. (2017) proposed a framework for automatic curriculum learning and examined training task selection. While their work inspired our target task, they did not investigate which source tasks are more beneficial for learning the target task.

Reward Shaping provides the agent with an *additional* reward to improve its performance. This reward is provided by the designer, not from the environment, and estimates how well an agent is currently achieving its task. Shaping aims to decrease exploration time, i.e., the time the agent explores suboptimal actions, which more efficiently explores larger action spaces. Shaping functions are often designed to be task specific (e.g., Ng et al. (1999) use relative distance to shape the reward). Recent work by Florensa et al. (2017) created source tasks increasingly further from the target and examined shaping for a curriculum within a single task. In that study, shaping provided no improvement.

Experiments

Our DQN, inspired by Mnih et al. (2015), consists of 3 convoluntional layers sized 32x96, 13x45, and 2x10 respectively. The input image to our network is scaled to 640x480. A frameskip of 5 allows performing a single action over many frames and speeds up training. In addition, the DQN uses experience replay of the past 60 experienced episodes. From this batch of episodes, a separate target network is trained on it to generate target Q values, ultimately providing stability to the network.

We performed a complete factorial experiment where we varied the training regime (T), color (C), shaping (S), and testing room type (R). The agent moves constantly forward at max speed and can the turn at a speed of 0, 0.3, or -0.3 of its maximum turning speed. We trained networks according to four curricula: $\{$AA, AB, BA, BB$\}$, where A (B) denotes the top-left (bottom-left) room in Figure 1. Shaping, when enabled, decreased the reward with the L1 distance to the goal. Coloring, when enabled, consisted of using blue squares shown in Figure 1. This generated 16 networks (i.e., 4 training regimes × color on/off × shaping on/off).

We tested the networks on four two-room variations of the lower-right room of Figure 1, differing by the color: (a) no color, (b) color around the target, (c) color around the threshold, and (d) color around both. We created 30 rooms of each type with the start and target position randomized. We ran each of the 16 networks on these 120 rooms and collected the final distance to target, the time taken to reach the target (or time out at 10 seconds), the cumulative reward (without shaping), and the total number of actions.

Results show that the training regime and color hinting most impact the agent's ability to reach the target; shaping has less impact. We discuss the the final distance to the target analysis using a *factorial* ANOVA – which performs simultaneous pairwise tests between the factors – to highlight the factors that most impact performance. Results for time, reward, and number of actions follow a similar pattern, though space limits prevent their inclusion.

Table 1 summarizes the factorial ANOVA for distance and is ordered by the increasing p-value for $p < 0.05$ (right column). Factors that significantly impact the distance, evidenced by $p \ll 0.01$, are color with training regime (C:T), training regime alone (T), and training regime with room type (T:R). Shaping (S) appears in later rows of the distance table indicating its effect is less significant; however, shaping does appear to have more impact for time and actions. Higher values in the "SumSq" column indicate greater contribution of the factor(s) to the variability in performance.

Factor	Df	SumSq	MeanSq	Fval	Pr(>F)
C:T	3	598	199.4	19.71	1.41e-12
T	3	364	121.3	12.00	8.87e-08
T:R	9	392	43.6	4.31	1.44e-05
C:S:T	3	107	35.7	3.53	0.0144
R	3	102	33.9	3.36	0.0182
S:T	3	86	28.7	2.84	0.0370
C	1	43	43.3	4.28	0.0388
C:T:R	9	176	19.6	1.94	0.0432
S:R	3	80	26.7	2.64	0.0482
Residuals	1856	18771	10.1		

Table 1: Multi-factor ANOVA on final distance from target for Color (C), Shaping (S), Training Regime (T), and Room (R). Only rows with $p < 0.05$ are shown.

Summary and Future Work

We examined the role of reward shaping, curricula, and visual hints in a Deep RL agent and found that the combination of curricula and hints had the most impact followed by curricula alone. This suggests that curricula design plays the biggest role for our study. Our methodology establishes an evaluation protocol for eventually understanding when and why these methods are applicable. Future work will generalize these results to a robotics domain, more complex tasks, and more sophisticated curricula, hints, and reward shaping.

Acknowledgments

The authors thank NRL for funding this research and for reviewer comments that helped improve the presentation.

References

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. *Proc. ICML* 41–48.

Florensa, C.; Held, D.; Wulfmeier, M.; and Abbeel, P. 2017. Reverse curriculum generation for reinforcement learning. *arXiv:1707.05300*.

Johnson, M.; Hofmann, K.; Hutton, T.; and Bignell, D. 2016. The Malmo platform for artificial intelligence experimentation. In *Proc. IJCAI*, 4246–4247.

Matiisen, T.; Oliver, A.; Cohen, T.; and Schulman, J. 2017. Teacher-student curriculum learning. *arXiv:1707.00183*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.

Narvekar, S.; Sinapov, J.; Leonetti, M.; and Stone, P. 2016. Source task creation for curriculum learning. In *Proc. AAMAS*, 566–574.

Ng, A.; Harada, D.; and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proc. ICML*, 278–287.